

**AN ARTIFICIAL NEURAL NETWORK  
FOR PATTERN CLASSIFICATION AND VISUALISATION**

**SANDRA ONG PI YIN**

This project is submitted in partial fulfilment of the requirements for a  
Bachelor of Science with Honours  
(Cognitive Science)

Faculty of Cognitive Sciences and Human Development  
UNIVERSITI MALAYSIA SARAWAK  
2010

## BORANG PENGESAHAN STATUS TESIS

**Gred:**JUDUL : \_\_\_\_\_  
\_\_\_\_\_

SESI PENGAJIAN : \_\_\_\_\_

Saya \_\_\_\_\_  
(HURUF BESAR)

mengaku membenarkan tesis \* ini disimpan di Pusat Khidmat Maklumat Akademik, Universiti Malaysia Sarawak dengan syarat-syarat kegunaan seperti berikut:

1. Tesis adalah hakmilik Universiti Malaysia Sarawak.
2. Pusat Khidmat Maklumat Akademik, Universiti Malaysia Sarawak dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Pusat Khidmat Maklumat Akademik, Universiti Malaysia Sarawak dibenarkan membuat pendigitan untuk membangunkan Pangkalan Data Kandungan Tempatan.
4. Pusat Khidmat Maklumat Akademik, Universiti Malaysia Sarawak dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.

\*\* sila tandakan ( ✓ )

☐

SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan seperti termaktub di dalam AKTA RAHSIA RASMI 1972)

☐

TERHAD

(Mengandungi maklumat Terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

☐

TIDAK TERHAD

\_\_\_\_\_  
(TANDATANGAN PENULIS)\_\_\_\_\_  
(TANDATANGAN PENYELIA)

Alamat Tetap:

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

Tarikh : \_\_\_\_\_

Tarikh: \_\_\_\_\_

**Catatan:**

\* Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah, Sarjana dan Sarjana Muda

\*Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai TERHAD.

The project entitled ‘An Artificial Neural Network for Pattern Classification and Visualisation’ was prepared by Sandra Ong Pi Yin and submitted to the Faculty of Cognitive Sciences and Human Development in partial fulfillment of the requirements for a Bachelor of Science with Honours (Cognitive Science).

Received for examination by:

-----  
(Assoc. Prof. Dr. Teh Chee Siong)

Date:

-----

<b>Grade</b>
--------------

### Statement of Originality

The work described in this Final Year Project, entitled  
**“An Artificial Neural Network for Pattern Classification and Visualisation”**  
is to the best of the author’s knowledge that of the author except  
where due reference is made.

---

(Date submitted)

---

(Student’s signature)  
Sandra Ong Pi Yin  
19809

## **ACKNOWLEDGMENT**

First and foremost, I would like to thank the person behind the motivation for completing this final year project, my supervisor, Dr. Teh Chee Siong. Dr. Teh has been very patient and kind in supervising my every step in order for me to successfully complete this task. I am eternally grateful for his guidance and support. I would also like to extend thanks to Mr. Yii Ming Leong, my laboratory demonstrator for Matlab software. Matlab is the main software used for this project. Mr. Yii has been willing to spare some of his time in teaching me the ins and outs of the software while sharing with me strategies to tackle the technical problems I faced during the completion of the project. Lastly, I'd like to acknowledge my friends and family. I really appreciated the constant support all of you provided throughout the most stressful time of the year. Thank you.

## TABLE OF CONTENTS

<b>Acknowledgment</b>	<b>iv</b>
<b>Table of Contents</b>	<b>v</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>xi</b>
<b>Abstract</b>	<b>xviii</b>
<b><i>Abstrak</i></b>	<b>xix</b>

### CHAPTER 1 - INTRODUCTION

1.0	Introduction	1
1.1	Background of Study	1
1.2	Artificial Neural Network	2
1.3	Data Classification	3
1.4	Data Visualisation	4
1.5	Research Motivation	4
1.6	Objectives	5
	1.6.1 General Objective	5
	1.6.2 Specific Objective	5
1.7	Scope of Project	6

### CHAPTER 2 - LITERATURE REVIEW

2.0	Introduction	7
<b>2.0</b>	<b>Pattern Recognition</b>	<b>7</b>
<b>2.1</b>	<b>Artificial Neural Network (ANN)</b>	<b>8</b>
<b>2.2</b>	<b>Statistical Pattern Recognition</b>	<b>10</b>
	2.3.1 Supervised Classification	10
	2.3.2 Unsupervised Classification	10
<b>2.3</b>	<b>Self-Organizing Map (SOM)</b>	<b>11</b>

## **CHAPTER 3 - SOM TOOLBOX AND INTRODUCTION TO EXPERIMENT**

3.0	Introduction	13
3.1	Description of the SOM Map	13
3.2	Mathematical Equations for SOM Training	15
3.3	Toolbox Details	16
3.3.1	Data Preprocessing	17
3.3.2	Data Initialization and Training	18
3.3.3	Others	19
3.4	Introduction to the experiment	21
3.5	Simulation Set One- 2D Gaussian Data set	23
3.5.1	Experiment I- Distance between each Gaussian	24
3.5.2	Experiment II- Data population	26
3.5.3	Experiment III- Number of classes	27
3.6	Simulation Set Two- 3D Gaussian Data set	28
3.6.1	Experiment I- Distance between each Gaussian	29
3.6.2	Experiment II- Data population	31
3.6.3	Experiment III- Number of classes	32
3.7	Data Pre-processing	33
3.8	Training Gaussian Data set using SOM	34
3.9	Quality of SOM training	38
3.10	User Interface of Classifier	39
3.11	Conclusion	40



## **CHAPTER 4 - CLASSIFICATION AND VISUALISATION**

4.0	Introduction	41
4.1	Constructing a Classifier	41
4.1.1	The Labeling Process	42
4.1.2	Visualisation of Classifier	44
4.2	Performance Analysis of SOM Classification	46
4.2.1	Classification Accuracy for 2D Gaussian	46
4.2.1.1	Analysis I- Different Distance	47
4.2.1.2	Analysis II- Increasing number of classes (3 classes)	50
4.2.1.3	Analysis III- Increasing Data Population	52
4.2.2	Classification Accuracy for 3D Gaussian	55
4.2.2.1	Analysis I- Different in Distance	56
4.2.2.2	Analysis II- Increasing number of classes (3 classes)	58
4.2.2.3	Analysis III- Increasing Data Population	60
4.3	Conclusion	63

## **CHAPTER 5 - HIGH DIMENSIONAL DATA SET CLASSIFICATION AND VISUALISATION**

5.0	Introduction	64
5.1	High Dimensional Data set	64
5.2	Classification Accuracy of Real World Data Set	65
5.3	Parametric Study	68
5.3.1	Effects of Map size	68
5.3.2	Effects of Learning Rate	68
5.3.3	Effects of Training length	68
5.4	Results Bench Marking	69
5.5	Conclusion	71

## **CHAPTER 6 - CONCLUSION**

6.0	Introduction	72
6.1	Strengths of the classifier	72
6.2	Weakness of the classifier	73
6.3	Recommendation for future work	73
6.4	Conclusion	74

<b>REFERENCE</b>	75
------------------	----

## LIST OF TABLES

### Table 3.1

Various type of Gaussian data set	22
-----------------------------------	----

### Table 4.1

2D low population Gaussian data set (with classes separated far) classification performance	47
--	----

### Table 4.2

2D low population Gaussian data set (with classes separated near) classification performance	48
---	----

### Table 4.3

2D low population Gaussian data set (with classes overlapping) classification performance	49
--	----

### Table 4.4

Classification performance using data set with 3 Classes	52
--	----

### Table 4.5

Classification performance using 2D data set with both low and high populated data set with 2 and 3 classes	54
--	----

### Table 4.6

3D low population Gaussian data set (with 2 classes) classification performance	57
--	----

### Table 4.7

3D low population Gaussian data set (with 3 classes) classification performance	59
--	----

**Table 4.8**

Classification performance using 2D data set with both low and high populated data set with 2 and 3 classes 62

**Table 5.1**

The parameters of the data set consist of factors that contributes to the possibilities of being a diabetic 65

**Table 5.2**

Table of classification performance for Pima Indians Diabetes Data set using SOM classifier 65

**Table 5.3**

Bench marking with similar study research 70

## LIST OF FIGURES

### **Figure 2.1**

McCulloch-Pitts neuron. 9

### **Figure 2.2**

Self-Organizing Map adopted from Ru and Horowitz (2007) 12

### **Figure 3.1**

(a) The data set used to train SOM (b) SOM map before training and,  
(c) Image result of SOM after training using a self-generated 2D Gaussian  
data set and the SOM toolbox 14

### **Figure 3.2**

The U-matrix on the left top corner and the component planes. 20

### **Figure 3.3**

The labels (names) for the nodes are shown on the right side of the  
Diagram 21

### **Figure 3.4**

A 2D Gaussian data set with 200 data for each class 24

### **Figure 3.5**

The two Gaussians overlapping each another 25

### **Figure 3.6**

The two Gaussians separated from each another 25

**Figure 3.7**

The two Gaussians separated further from each another 25

**Figure 3.8**

Low populated 2D Gaussian that consists of 200 data, 100 data point per class 26

**Figure 3.9**

High populated 2D Gaussian that consists of 2000 data, 1000 data points per class. However, class 2 (grey) have higher density compared to class 1 (black). 26

**Figure 3.10**

A 2D Gaussian data set with 3 classes distinctively indicated with the different greyscale colours (black, dark grey and light grey) 27

**Figure 3.11**

3D Gaussian data set with 600 data, 300 data each class. Class 2 (black data points) uses a relatively higher SD value compared to class 1 (grey data points). 29

**Figure 3.12**

3D Gaussian overlapping each another 29

**Figure 3.13**

3D Gaussian separated from each another 30

**Figure 3.14**

3D Gaussian separated further from each another 30

**Figure 3.15**

3D Gaussian data set that consist of 300 data, 150 data points each class 31

**Figure 3.16**

3D Gaussian data set that consist of 3000 data, 1500 data points each class 31

**Figure 3.17**

A 3D Gaussian data set that consist of 3 classes (red, black and grey)  
overlapping each another 32

**Figure 3.18**

A 3D Gaussian data set that consist of 3 classes (grey, dark grey and  
black) separated from each another 32

**Figure 3.19**

The separation of data set that consist of 400 data (a) 90 percent data,  
(b) 10 percent and, (c) 100 percent or all data 34

**Figure 3.20**

Randomly initialised SOM map that consist of 100 nodes and links for each  
node 35

**Figure 3.21**

An example of 2D data set that contains 200 data points fed to the  
classifier 35

**Figure 3.22**

SOM map structure at the final epoch of the training process, together  
displayed with the data set used to train the nodes. 37

**Figure 3.23**

Mean Quantization Error graph show a decrease in error 38

**Figure 3.24**

Topology Error 38

**Figure 3.25**

The GUI for the classifier created for this project 39

**Figure 4.1**

The process of labeling the codebook 42

**Figure 4.2**

A 2 dimensional data set with additional 3 arrays (columns) behind to stored the counts for class 1, class 2 and the finalized decision of class label respectively. 43

**Figure 4.3**

Codebook grid form shows the position of the codebook with class information: Class 1 (green), class 2 (red), and empty (black) 45

**Figure 4.4**

Hexagonal grid that displays bar charts that indicate the count of “class 1” and “class 2” for each subunit. The hexagonal charts without any bars (blue) are considered not active. The background colour of each subunit shows the class classification. 46

**Figure 4.5**

(a) 2D Gaussian data set that is separated far (b) Visualisation of classification for the data set 47



**Figure 4.6**

(a) 2D Gaussian data set that is separated near (b) Visualisation of classification for the data set 48

**Figure 4.7**

(a) 2D Gaussian data set that overlaps (b) Visualisation of classification for the data set 49

**Figure 4.8**

(a) A 2D Gaussian data set with 3 classes that are separated far and (b) its visualisation of classification for the data set 50

**Figure 4.9**

(a) A 2D Gaussian data set with 3 classes that are separated near and (b) its visualisation of classification for the data set 50

**Figure 4.10**

(a) A 2D Gaussian data set with 3 classes that overlaps and (b) its visualisation of classification for the data set 51

**Figure 4.11**

(a) A 2D Gaussian data set (2 classes) with low data population and (b) its visualisation of classification 52

**Figure 4.12**

(a) A 2D Gaussian data set (2 classes) with high data population and (b) its visualisation of classification 53

**Figure 4.13**

- (a) A 2D Gaussian data set (3 classes) with low data population and  
(b) its visualisation of classification 53

**Figure 4.14**

- (a) A 2D Gaussian data set (3 classes) with high data population and  
(b) its visualisation of classification 54

**Figure 4.15**

- (a) 3D Gaussian data set that is separated far and (b) Visualisation of  
classification for the data set 56

**Figure 4.16**

- (a) 3D Gaussian data set that is separated near and (b) Visualisation of  
classification for the data set 56

**Figure 4.17**

- (a) 3D Gaussian data set that overlaps (extreme) and (b) Visualisation of  
classification for the data set 57

**Figure 4.18**

- (a) 3D Gaussian data set with 2 classes and (b) Visualisation of  
classification for the data set 58

**Figure 4.19**

- (a) 3D Gaussian data set with 3 classes and (b) Visualisation of  
classification for the data set 58

**Figure 4.20**

- 3D Gaussian with 3 classes that is extremely overlapped while being

clustered together 60

**Figure 4.21**

(a) A low populated 3D Gaussian data set and (b) Visualisation of classification for the data set 60

**Figure 4.22**

(a) A high populated 3D Gaussian data set and (b) Visualisation of classification for the data set 61

**Figure 4.23**

(a) A low populated 3D Gaussian data set and (b) Visualisation of classification for the data set 61

**Figure 4.24**

(a) A high populated 3D Gaussian data set and (b) Visualisation of classification for the data set 62

**Figure 5.1**

Visualisation of classification performance of the SOM map using the Pima Indians Diabetes data set 67

## **ABSTRACT**

### ***AN ARTIFICIAL NEURAL NETWORK FOR PATTERN CLASSIFICATION AND VISUALISATION***

Sandra Ong Pi Yin

The industrial revolution and the birth of computers has led to a deeper exploration of Artificial Neural Network (ANN), where scientist tries to emulate the biological neural network. Today, ANN has proven to be able to imitate the human neural network and perform task such as solving real world problems. This study aims to explore in detail an ANN model that is able to perform the task of pattern classification and visualisation, and as well as to evaluate the performance of this model. This proposed model is know as the Self-Organizing Map (SOM) or Kohonen's Map. In order to determine it's accuracy, the SOM classifier is tested using a few simulated Gaussian data sets and a real world data set, the Pima Indians Diabetes data set. The experiments conducted showed that SOM classifier is able to perform the task of classification and visualisation. However, the classifier obtained a non-impressive accuracy of 73.16% in classifying the real world problem. This results indicate that SOM is not reliable compared to other classification applications in literature and can be enhanced in terms of it's performance.

## **ABSTRAK**

### ***AN ARTIFICIAL NEURAL NETWORK FOR PATTERN CLASSIFICATION AND VISUALISATION***

Sandra Ong Pi Yin

*Revolusi industri dan penciptaan komputer telah membawa kepada pendalaman dalam pemahaman tentang Rangkaian Neural Buatan. Ahli Sains telah membuktikan bahawa ANN dapat meniru rangkaian neural manusia dan mampu melaksanakan tugas seperti manusia. Projek ini bertujuan untuk mengkaji model ANN yang dapat menjalankan tugas klasifikasi pola serta menilai ketepatan pencapaian model tersebut. Model yang dicadangkan dikenali sebagai Self-Organizing Map (SOM) ataupun Kohonen's Map. SOM diuji dengan menggunakan beberapa set data simulasi Gaussian dan data Pima Indians Diabetes untuk menentukan ketepatan klasifikasinya. Hasil ujikaji menunjukkan SOM mampu menjalankan tugas klasifikasi di samping memberi gambaran pola set data yang digunakan. Akan tetapi, pengklasifikasian mencapai ketepatan hanya sebanyak 73.16% bagi data Pima Indians Diabetes. Dapatan ini menunjukkan bahawa prestasi SOM tidak memuaskan jika berbanding dengan aplikasi pengklasifikasian yang lain.*

## **CHAPTER 1**

### ***INTRODUCTION***

#### **1.0 Introduction**

This chapter briefly introduces the background of the project, the motivation behind the project as well as the objectives and scope of the project.

#### **1.8 Background of Study**

Technology advancement began during the industrial revolution beginning in the 1900s. This advancement led to the birth of computers later on in the 1950's. However, the idea of emulating the biological neural networks began before computers came about. In the 1940's, scientist and researchers have already begin exploring the world of Artificial Neural Network (ANN) or also known as Neural Networks (NN). Many theories have formed throughout the years and ANN continued to bloom until the period between 1969 and 1981. The sudden downturn towards ANN was due to media spread of fiction stories regarding the ability of neural network accomplishing many tasks. Therefore, it was not welcomed by the society for they fear robots would trump mankind and dominate the world. However, in the 1982, the idea of neural networks was once again convinced when John Hopfield of Caltech stated that the approach of

Artificial Intelligence should not completely focus on the means of imitating the human neural network but to create machines that are able to solve dynamic problems (Oracle, 2008). Today, ANN has proven to be beneficial to society in performing tasks. Countries around the world have pursued and invested a great deal in the research and development of ANN.

## **1.9 Artificial Neural Network**

Artificial Neural Network (ANN) is the attempt in mimicking the patterns of the human mind (Botts, 1998). According to the Encyclopedia of Educational Technology, the idea of the artificial neural network emerged back in the early 1940's when neuroscientist, Warren McCulloch and mathematician, Walter Pitts hypothesized that "neurons in our nervous system could be regarded as a device for manipulating binary numbers, the computer" (1998). By joining venture, they designed and built a primitive artificial neural network by using electric circuits. Their work has led to the infamous McCulloch-Pitts Theory of Formal Neural Networks (Oracle, 2008).

The next major development in neural network technology was in the year 1959, when Bernard Wildro and Marcian Hoff of the Stanford University USA developed the first neural network that can be applied to real world problems. It is known as the Adaptive Linear Elements (ADALINE) and Multiple Adaptive Linear Elements (MADELINE) (Oracle, 2008).

Advancement in ANN has allowed an increasing number of complex real world problems to be solved. A few of many ANN applications in real life includes classification which consist of pattern and sequence recognition, sequential decision making; data processing which consist of filtering, and clustering; and many more (Wikipedia, 2009).

## **1.10 Data Classification**

What is data? Data are factual information organized for analysis or used for decision making purposes (Answers, 2010). In the field of computers, data refers to a collection of numbers and symbols, represented in a form suitable to be processed by computers (Himberg, 2004). These data might be obtained by scientific experiments, large databases, or collections of digital documents (Himberg, 2004).

Classification infers the act of classifying. It is a fundamental process that is carried out daily by human beings. Human beings are programmed to have the ability to categorize things so it would be easier to identify or distinguish the differences or the similarity between the groups of things (Modell, 1992). According to Modell (1992), “to classify is to organize or arrange according to class or category”. These categories contain data or information that has at least one similarity between them. For example, the common attribute for the class ‘fruits’ is that it contains seeds. Therefore, ‘apple’ would be classified as a fruit instead of vegetable because it fulfills the required attribute. Data classification may also be done according to the importance of that data or how frequent it would be accessed, data content, data size, or even categorized based on who uses the data (TechTarget, 2007). These common attributes or similarity in characteristics or features is known as parameters.

Today, there are many neural network models that have been recognized to perform the task of data classification. These include the simple perceptron model, the Adaptive Linear Element (ADALINE), Multi-Layer Perceptron (MLP), Self-Organizing Map (SOM), the Learning Vector Quantization (LVQ) and more.